Enhancing Bioactivity Prediction via Spatial Emptiness Representation of Protein-ligand Complex and Union of Multiple Pockets

Zhiyuan Zhou *
School of Biological Sciences
Nanyang Technological University
Singapore 637551
zhou0457@e.ntu.edu.sg

Yueming Yin *
College of Computing and Data Science
Nanyang Technological University
Singapore 639798
yueming.yin@ntu.edu.sg

Yiming Yang
School of Biological Sciences
Nanyang Technological University
Singapore 637551
yiming014@e.ntu.edu.sg

Yuguang Mu
School of Biological Sciences
Nanyang Technological University
Singapore 637551
ygmu@ntu.edu.sg

Hoi-Yeung Li †
School of Biological Sciences
Nanyang Technological University
Singapore 637551
hyli@ntu.edu.sg

Adams Wai-Kin Kong †
College of Computing and Data Science
Nanyang Technological University
Singapore 639798
adamskong@ntu.edu.sg

Abstract

Predicting the bioactivity of candidate ligands remains a central challenge in drug discovery. Ligands and endogenous substrates often compete for the same binding sites on target proteins, and the extent to which a ligand can modulate protein function depends not only on its binding but also on how effectively it occupies the relevant pocket. However, most existing methods focus narrowly on local interactions within protein–ligand complexes and neglect spatial emptiness—the unoccupied regions within the binding site that may permit endogenous molecules to engage or interfere. Such unfilled space can diminish the ligand's functional impact, regardless of binding affinity. To overcome this key limitation in protein–ligand modeling, we propose **LigoSpace**, a novel method integrating three core components. LigoSpace introduces **GeoREC** (Geometric Representation of Spatial Emptiness in Complexes) to quantify atomic-level empty space and **UnionPocket** to unify multiple protein pockets, providing a global view of binding sites. Additionally, LigoSpace employs a **pairwise loss** instead of commonly used MSE loss, to better capture relative relationships critical for drug discovery. Extensive experiments on multiple datasets with diverse bioactivity types demonstrate that LigoSpace significantly improves performance when integrated into state-of-the-art models, highlighting the effectiveness of its novel components.

*Equal contribution, may cite either first.
†Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

https://doi.org/10.52202/085713-3543

1 Introduction

Bioactivity is the measurable effect a compound exerts on a biological system, usually by modulating the functional activity of a specific molecular target (e.g., a protein or nucleic acid). It can be quantified experimentally through enzymatic kinetics ($V_{\max}$, K_{m}, $k_{\mathrm{cat}}/K_{\mathrm{m}}$) or functional and binding assays that yield parameters such as EC_{50}, IC_{50}, $E_{\max}$ and K_{i}) [MarÉchal, 2011, Mendez et al., 2019]. Accurate measurement—and increasingly, in-silico prediction—of these effects allows researchers in drug discovery, chemical biology, and pharmacology to identify the most promising compounds early, reducing the cost and duration of large-scale experimental screens [Kapetanovic, 2008]. With the growing size of chemical libraries and biological data, there is an increasing demand for efficient and accurate computational methods to predict bioactivity. Developing robust bioactivity prediction methods has therefore become a key focus in both academic research and industrial applications [Ramsundar et al., 2019].

Recently, machine learning approaches have shown great potential in modeling biological molecules and their interactions, dramatically accelerating the drug discovery process [Rifaioglu et al., 2019, Chen et al., 2018]. There have also been some machine learning methods for bioactivity prediction, aiming at early-stage screening of the potential active molecules [Prajapati et al., 2025, Ishfaq et al., 2022, Cáceres et al., 2020, Lee et al., 2019]. Conventional approaches rely on the intrinsic properties of small molecules as features to model ligands, aiming to uncover the relationship between molecular structure and bioactivity. A traditional strategy is to use quantitative structure–activity relationship (QSAR) models [Cherkasov et al., 2014, Tropsha et al., 2024], which primarily utilize molecular descriptors derived from the physicochemical and structural properties of small molecules. While this approach is reasonable and widely adopted, it may overlook the fact that bioactivity fundamentally arises from specific biological interactions, which are critical for thoroughly modeling and understanding bioactivity. A compound cannot function biologically without its target and environment. Furthermore, unlike other single-target prediction tasks, bioactivity is often evaluated using multiple measurement types, such as IC_{50}, EC_{50}, K_{d}, and K_{i}, making the task more challenging. In response to these challenges, recent studies have shifted toward developing models based on curated pocket-ligand interaction datasets that provide more comprehensive information, including abundant bound conformations with proteins and multiple bioactivity measurements [Yin et al., 2024, Huang et al., 2025]. Among these methods designed for protein-ligand interaction, graph neural networks (GNNs) have become particularly prevalent, as graph-based representations offer a natural and powerful way to capture molecular properties and relational structures [Zhang et al., 2023, Wang et al., 2024, Mastropietro et al., 2023, Li et al., 2021].

While recent studies increasingly focus on bioactivity prediction using protein information and constructing protein-ligand interaction graphs, they often concentrate on local binding region but overlook the global geometric context of the entire pocket, missing crucial information for accurate bioactivity modeling. Geometric features are known to be important for representing binding status and have shown effectiveness in modeling molecular interactions [Wei et al., 2024, Han et al., 2024, Song et al., 2024, Zhang et al., 2022]. For instance, geometric representations in E(3)-equivariant GNNs have demonstrated strong performance in capturing local spatial relationships on protein surfaces, significantly improving binding site prediction [Wei et al., 2024]. However, most existing approaches emphasize detailed descriptions of local binding region and structural complementarity between interacting molecules. None of the previous work considers pocket–ligand interactions from the perspective of spatial emptiness within the docked conformation. This aspect, which remains underexplored, holds practical importance and corresponds closely with the principle of competitive inhibition [Eun, 1996, Baici, 2015], where drug molecules bind to the active site of a protein and block access for endogenous substrates (illustrated in Figure 1). For example, protein kinase inhibitors bind to the enzyme's ATP-binding pocket, blocking ATP access and thereby suppressing kinase activity. Previous methods tend to focus on how well the ligand fits the pocket, but often overlook the inverse questions: once binding occurs, how much empty space remains, and to what extent does the ligand occupy the pocket?

Furthermore, current methods for bioactivity prediction typically focus on the local binding pocket in complex with a ligand, which limits access to global pocket information that is biologically significant. Additionally, in many prior works, each pocket–ligand pair is treated as an independent data point [Lee et al., 2019, Cáceres et al., 2020], overlooking the realistic scenario in which multiple potential drug compounds may interact with different pockets of the same protein. Recent findings underscore

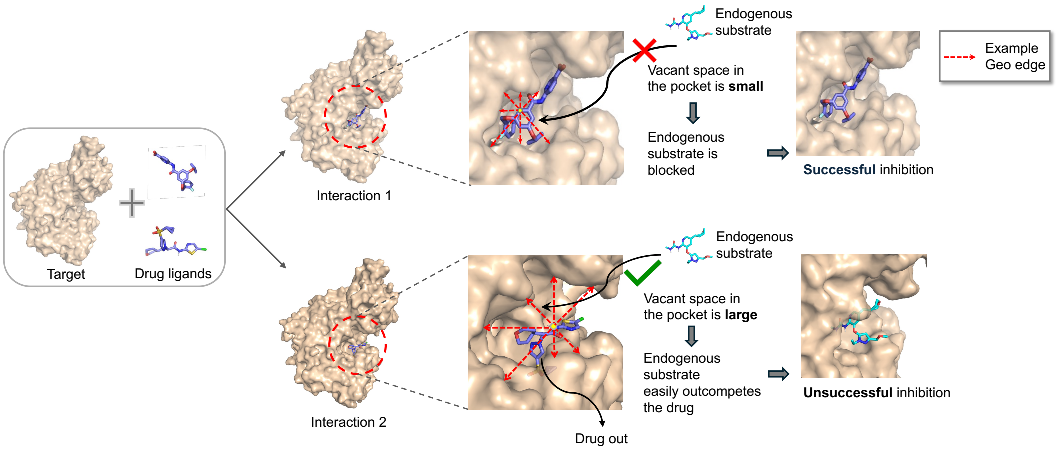

Figure 1: An example of competitive inhibition.

the importance of evaluating bioactivity across different ligands acting on a common protein, enabling more meaningful and fair comparisons. In response, new methods and datasets [Yin et al., 2024, Huang et al., 2025] have been developed to incorporate multiple ligands per protein and to perform evaluation at the protein level, allowing for more robust and context-aware assessments. However, even with a fixed protein, these approaches often involve distinct binding sites within the same protein, introducing variability that may compromise the consistency of bioactivity predictions.

Additionally, a major distinction between drug discovery and other regression tasks lies in the importance of prediction order. The most common training loss, mean-square error (MSE), is symmetric and treats each data point independently. For example, given two ground-truth bioactivities, $y_1 = 2$ and $y_2 = 1$, it gives the same loss value for the predictions $(\hat{y}_1 = 1.5, \hat{y}_2 = 1.5)$ and $(\hat{y}_1 = 1.5, \hat{y}_2 = 0.5)$. The order of the former is wrong, but the order of the latter aligns with the ground truths.

To address these issues, we propose LigoSpace, which integrates three key modules for enhanced bioactivity prediction. We make the following key contributions:

- We propose **Geo**metric **R**epresentation of spatial **E**mptiness in protein-ligand **C**omplex (GeoREC) to describe the surroundings of each node in the protein-ligand graph representation. This surrounding space quantifies the emptiness within the docking conformation, incorporating a broader global perspective into the overall interaction graph representation. Furthermore, this concept of emptiness holds practical significance, as it plays a pivotal role in assessing whether the pocket space is sufficiently tight to hinder the entry of other small molecules into the binding site area, thereby contributing to a more accurate prediction of the compound's bioactivity.
- We introduce the concept of the Union-Pocket (Figure 3) into the bioactivity prediction task to ensure that the entire pocket space can be accessible, facilitating GeoREC to measure the empty space inside. Moreover, it provides a broader structural context during the GNN processing, contributing to more effective global information integration. Additionally, when evaluating different ligands binding to the same protein, keeping the protein information consistent encourages the model to focus on the ligand, the only changing part, enhancing the ability of the model to capture subtle conformational differences within a consistent protein context.
- We also incorporate a pairwise loss alongside the conventional mean squared error (MSE) loss as the final training objective for bioactivity prediction. This combined loss encourages the model to preserve the relative order among samples by emphasizing pairwise differences in their bioactivity labels.
- Experimental results demonstrate that LigoSpace leads to substantial improvements in bioactivity prediction across multiple datasets, label types, and GNN architectures. In

https://doi.org/10.52202/085713-3543

particular, the addition of GeoREC results in significant performance gains, highlighting the critical role of geometric information in accurately modeling protein-ligand interactions.

2 Related work and preliminaries

2.1 Bioactivity prediction

Predicting bioactivities is a crucial step in drug discovery, as it plays a central role in enabling efficient and accurate drug screening [Ishfaq et al., 2022, Yin et al., 2024]. Early empirical approaches [Gohlke et al., 2000, Wang et al., 2002] rely on manually designed scoring functions tailored for specific tasks, requiring substantial expert knowledge to encode underlying biochemical interactions. Subsequently, statistical and traditional machine learning methods [Ballester and Mitchell, 2010, Kinnings et al., 2011] are developed to model bioactivity in a data-driven manner by extracting protein and ligand features and applying classical regression algorithms. However, these methods heavily depend on the quality of handcrafted features and often suffer from limited generalizability when applied to larger or more diverse datasets. With the rising popularity of deep learning in recent years, sequence-based bioactivity prediction methods have emerged, which directly extract features from the SMILES (Simplified Molecular Input Line Entry System) strings of ligands and the amino acid sequences of proteins, bypassing the need for manual feature engineering. For example, DeepDTA [Öztürk et al., 2018] employs two independent convolutional neural networks to capture latent representations from ligand SMILES and protein sequences. Building upon this, MT-DTI [Shin et al., 2019] incorporates attention mechanisms to enhance the model's interpretability and capture long-range dependencies.

Inspired by the effectiveness of graph neural networks (GNNs) in handling structured data, recent efforts [Vefghi et al., 2025] have extended bioactivity prediction to incorporate molecular graph representations. These approaches consider 3D structural information of protein-ligand complexes to varying degrees. GIGN [Yang et al., 2023] introduces a heterogeneous interaction layer that integrates both covalent and non-covalent interactions into the message-passing framework, thereby facilitating more effective node representation learning. DTIGN [Yin et al., 2024] enhances upon this by incorporating intra-molecular bond features and explicitly modeling non-covalent interactions, including Coulomb and London dispersion forces, and further combining molecular docking with self-attention mechanisms to capture information from diverse binding poses. However, existing models represent protein-ligand complexes as purely topological graphs, thereby underutilizing rich biomolecular structural information and neglecting the spatial interactions between proteins and ligands. To address these limitations, we propose incorporating comprehensive spatial information to better capture the geometric nature of molecular interactions.

2.2 Task definition

We consider the task of predicting continuous-valued bioactivity based on a set of possible binding poses of a ligand across multiple available protein pockets. Let $\mathcal{L}$ denote the set of ligands and $\mathcal{P}$ the set of protein pockets. Each ligand $l \in \mathcal{L}$ is associated with a set of N_l possible binding poses:

$$\mathcal{X}_l = \{x_{l,1}, x_{l,2}, \dots, x_{l,N_l}\}, \quad x_{l,i} \in \mathbb{R}^d, \tag{1}$$

where $x_{l,i}$ represents the i-th pose of the ligand l associated with one of the available protein pockets, and d is the dimensionality of the input representation (e.g., 3D coordinates, graph embeddings, or voxel grids). Let $\mathcal{P}_l \subseteq \mathcal{P}$ be the subset of candidate pockets associated with ligand l. Let $y_l \in \mathbb{R}$ denote the experimentally measured bioactivity value of ligand l (e.g., EC_{50}, IC_{50}), which may not be pocket-specific. The objective is to learn a predictive model f such that:

$$\hat{y}_l = f(\mathcal{X}_l, \mathcal{P}_l), \tag{2}$$

which aggregates information from the poses across all candidate pockets to estimate $\hat{y}_l$. In practice, pose-level representations can be computed and then aggregated using a permutation-invariant function:

$$\hat{y}_l = \mathrm{Agg}\left(\{f_{\text{pose}}(x_{l,i}, p_{l,i})\}_{i=1}^{N_l}\right), \tag{3}$$

where $p_{l,i} \in \mathcal{P}_l$ denotes the pocket corresponding to pose $x_{l,i}$, f_{pose} is a shared neural network applied to each pose-pocket pair, and Agg is a permutation-invariant aggregation function such as mean, sum, or attention-based pooling.

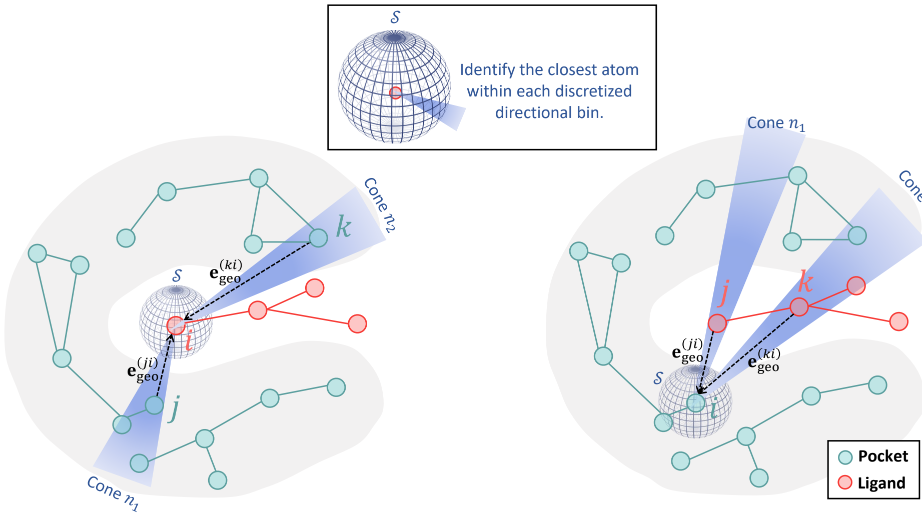

Figure 2: An example of the geometric edges defined in this paper.

3 Methods

3.1 Geometric representation of spatial emptiness in protein-ligand complex

To characterize the spatial configuration inside the pocket within a docked protein-ligand complex, we design **Geo**metric **R**epresentation of spatial **E**mptiness in protein-ligand **C**omplex (GeoREC) for each node that captures global information about the spatial emptiness of the structure. Let V_p and V_l denote the sets of nodes representing the pocket and the ligand, respectively, within a pocket-ligand complex. The complex graph is defined as $G_c = (V_c, E_c)$, where $V_c = V_p \cup V_l$, and E_c denotes the set of all types of the connection within the complex, usually corresponding to physicochemical interactions such as covalent bonds, hydrogen bonds, van der Waals contacts and so on.

For each node $v_i \in V_c$, we define a local spherical coordinate system centered at v_i, and uniformly partition the surrounding space into S cones. This spatial decomposition enables the extraction of relative positional features. For each cone $s \in \{1, 2, \ldots, S\}$, we collect a set of pocket-ligand nodes located within it and denote it by $V_{\text{zone}}^{(s)}$. Next, for each cone s whose corresponding node set $V_{\text{zone}}^{(s)}$ is non-empty, we identify the node closest to v_i, and include it in the surrounding node set $V_{\text{surr}}^{(i)}$. This set provides a compact yet informative representation of the spatial emptiness around v_i. In this study, GeoREC is incorporated by introducing geometric edges: for each $v_j \in V_{\text{surr}}^{(i)}$, we add an edge (v_i, v_j), forming a set of geometric edges denoted as E_{geo}. An example of the geometric edges is shown in Figure 2. The enhanced graph is then defined as: $G'_c = (V_c, E_c, E_{\text{geo}})$, which incorporates both the original interaction edges and the newly introduced geometric edges. These geometric edges enhance the graph's capacity to perceive spatial emptiness, particularly within the pocket cavity, and also provide alternative pathways for global and local information flow beyond conventional chemical bonds.

3.2 Global interaction model with Union-Pocket

The structure-based bioactivity prediction task can be generally formulated as

$$\hat{y}_l = f'(x_l, p), \tag{4}$$

where $p \in \mathcal{P}$ is a candidate binding pocket, x_l denotes the representation of the ligand l, f' is a generic bioactivity prediction model, and $\hat{y}_l$ is the predicted bioactivity. In this setting, each data point corresponds to a local pocket-ligand pair, e.g., "Data point 1" or "Data point 2" in Figure 3A. Conventional approaches [Wallach et al., 2015, Tanebe and Ishida, 2021, Yang et al., 2023] predict bioactivity on a single local pocket-ligand pair for a single ligand. However, a given ligand may engage multiple local pockets on the protein, each potentially contributing differently to the overall bioactivity.

https://doi.org/10.52202/085713-3543

To better reflect the biological context, recent work [Yin et al., 2024] has shifted toward aggregating multiple local pocket-ligand pairs for the same ligand. The prediction task then becomes:

$$\hat{y}_l = f(\mathcal{X}_l, \mathcal{P}'_l), \tag{5}$$

where $\mathcal{P}'_l \subseteq \mathcal{P}$ is the subset of local pockets associated with the ligand poses $\mathcal{X}_l$ within the same protein. This setting reflects the biological scenario more closely, as a single protein may engage the same ligand through distinct local pockets. Furthermore, a set of distinct ligands, denoted as $\mathcal{L}$, may generate numerous fragmented local pockets, introducing variability that hinders the model from capturing a global view of how all ligand poses—regardless of whether they come from the same or different ligands—relate on the same protein.

To mitigate this issue, we propose the concept of the *Union-Pocket*, defined as the union of all atoms from the candidate pockets associated with a given protein. Let $\mathcal{P}'_l = \{p'_1, p'_2, \ldots, p'_k\}$ be the set of pockets associated with ligand l. For each $p' \in \mathcal{P}'_l$, let $A(p')$ denote the set of atoms comprising pocket p'. The Union-Pocket is then defined as:

$$A(p_{\text{union}}) = \bigcup_{l \in \mathcal{L}} \bigcup_{p' \in \mathcal{P}'_l} A(p'), \tag{6}$$

where $\mathcal{L}$ is the set of ligands docked to the given protein. The resulting p_{union} aggregates all atoms from all relevant pockets across ligands and thereby encompasses the full binding region probed by the ligand set $\mathcal{L}$. An example of the formation and data representation of the proposed Union-Pocket are shown in Figure 3B and 3C, respectively. Accordingly, the bioactivity prediction task for each protein can now be reformulated as:

$$\hat{y}_l = f_{\text{union}}(\mathcal{X}_l, p_{\text{union}}), \tag{7}$$

where f_{union} is the prediction model utilizing the Union-Pocket.

By replacing many local pocket representations with the Union-Pocket, a model can capture richer global interaction information while maintaining spatial consistency across ligands binding to the same protein. When integrated with GeoREC, this strategy improves the geometric representation's expressiveness and allows the model to more effectively distinguish fine-grained differences in ligand poses and spatial emptiness. The shared spatial context provided by the Union-Pocket is crucial for accurate and robust bioactivity prediction.

3.3 The pairwise loss

To better preserve relative order information during optimization, we introduce a pairwise loss [Zhu et al., 2024, Tynes et al., 2021] into the bioactivity prediction task. Although recent studies have explored diverse metrics such as Kendall Tau [Fernández-Llaneza et al., 2021], Top-K precision [Yin et al., 2021], and enrichment factors [Yin et al., 2022] for evaluation, few have improved the loss function directly into model training. Most existing AI-based methods for bioactivity prediction and other biological regression tasks still rely on MSE loss to train their models. However, it is ineffective for drug discovery because it considers each prediction independently, neglecting the relative orders of the samples.

Consider the conventional MSE loss, which is also called L_2 loss, used for training the bioactivity prediction model:

$$\mathcal{L}_2 = \frac{1}{N} \sum_{l=1}^{N} \left\| \hat{y}_l - y_l \right\|_2^2, \tag{8}$$

where N is the total number of training samples (protein-ligand pairs) in a batch, $\hat{y}_l$ is the predicted bioactivity score for the l-th ligand, y_l is the ground truth bioactivity label for the l-th ligand and $\|\cdot\|_2^2$ denotes the squared L_2 norm. Standard loss functions such as L_2 (MSE) and L_1 (MAE) simply measure the absolute numerical differences between predicted and true values, ignoring relative order. This limitation poses challenges when capturing correlations among samples. We exploit the pairwise loss as follows:

$$\mathcal{L}_{\text{pairwise}} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left\| (y_i - y_j) - (\hat{y}_i - \hat{y}_j) \right\|_2^2. \tag{9}$$

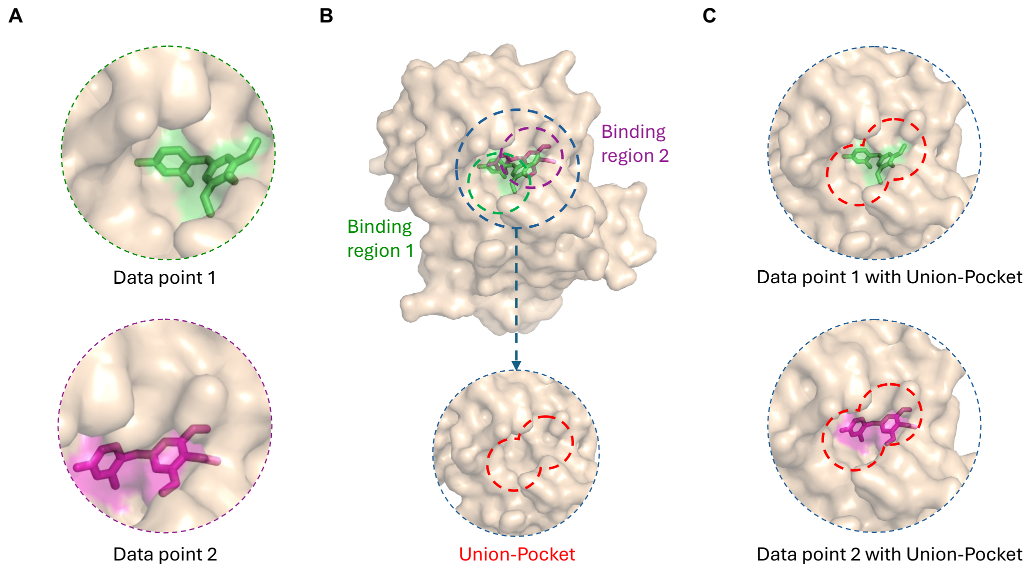

Figure 3: Union-Pocket. The protein is shown in wheat color. The stick models in green and magenta represent the same ligand, which binds to different regions of the protein. The green and magenta surfaces correspond to the surfaces of the ligand molecules in their respective binding conformations. **(A)** Current methods incorporate pocket-ligand interactions but focus on local regions, treating a small ligand-containing pocket as a single data point, thus losing global pocket information. **(B)** By taking the union of all the local binding regions within the protein, we get the Union-Pocket, indicated by the red dashed line. **(C)** New data points with our Union-Pocket, which keeps the complete global geometry of the binding region.

Here, N is the number of protein-ligand pairs in a batch. y_i and y_j are the ground truth bioactivity values for ligand l_i and l_j, respectively. $\hat{y}_i = f_{union}(\mathcal{X}_{l_i}, p_{union})$ and $\hat{y}_j = f_{union}(\mathcal{X}_{l_j}, p_{union})$ are the predicted bioactivity scores obtained from the model f_{union}. The pairwise difference $(y_i - y_j)$ reflects the relative bioactivity difference in ground truth, while $(\hat{y}_i - \hat{y}_j)$ is the corresponding prediction difference. The pairwise loss function employs an L_2 norm to penalize discrepancies in predicted relative differences $(\hat{y}_i - \hat{y}_j)$ compared to the true differences $(y_i - y_j)$. The normalization factor $\frac{2}{N(N-1)}$ computes the mean over all unique sample pairs, making the loss invariant to batch size. Ultimately, we adopt a hybrid loss function $\mathcal{L}_{\text{hybrid}}$ that combines both pointwise and pairwise losses:

$$\mathcal{L}_{\text{hybrid}} = \lambda_1 \cdot \mathcal{L}_2 + \lambda_2 \cdot \mathcal{L}_{\text{pairwise}}. \tag{10}$$

Here, $\mathcal{L}_2$ measures the pointwise error between predicted and true bioactivity values, and $\mathcal{L}_{\text{pairwise}}$ emphasizes the preservation of the correct difference between the predications. The coefficients λ_1 and λ_2 are hyperparameters that control the relative contribution of the two components in the overall objective.

4 Experiments

4.1 Datasets and testing settings

DTIGN dataset. We first adopt the dataset constructed in DTIGN [Yin et al., 2024]. This dataset is derived from ChEMBL [Gaulton et al., 2012], a widely used database in drug discovery that contains extensive information on chemical compounds, their biological activities, and related properties. The DTIGN dataset comprises 8 protein targets (I1, I2, I3, I4, I5, E1, E2, and E3), each considered as an individual subset containing several binding pockets. Within each subset, hundreds to thousands of ligands are associated with the protein, labeled by IC_{50} or EC_{50}, and each ligand is represented by multiple docking poses.

https://doi.org/10.52202/085713-3543

SIU dataset. We also validate our method on a large-scale docking dataset proposed in [Huang et al., 2025]. This dataset contains 201,458 and 56,485 small molecule–pocket pairs with K_{i} and K_{d} bioactivity labels, respectively, which complement the bioactivity categories in the DTIGN dataset. The K_{i} subset was generated by docking 34,881 unique small molecules to 4,317 protein conformations derived from 764 proteins, whereas the K_{d} subset was generated by docking 5,201 unique small molecules to 3,682 protein conformations derived from 584 proteins.

The dataset includes two variants, SIU-0.9 and SIU-0.6, sharing the same test set. The numeric suffix denotes the sequence similarity threshold between the training and test data. In this study, we use four subsets of the SIU dataset: SIU-0.9/K_{d}, SIU-0.9/K_{i}, SIU-0.6/K_{d}, and SIU-0.6/K_{i}.

Evaluation metrics. To ensure fair comparison with previous works, we adopt widely used evaluation metrics in bioactivity prediction, including root mean square error (RMSE), Pearson correlation coefficient (r), and Kendall's tau correlation coefficient (τ) for DTIGN dataset. We also compare mean absolute error (MAE) and Spearman correlation coefficient for SIU dataset. The details of the definition can be found in Appendix A.6.

Baseline models. To evaluate the effectiveness of the proposed representations and loss function, we adopt DTIGN [Yin et al., 2024], GIGN [Yang et al., 2023], and Graph Attention Network (GAT) [Veličković et al., 2017] as the baselines. DTIGN is one of the latest GNN-based methods designed for bioactivity prediction. It uses multiple local pocket-ligand complexes as input. GIGN is also a GNN-based method that incorporates 3D structures and physical interactions for predicting protein–ligand binding affinities. GAT is a foundational GNN model that uses attention to focus on relevant nodes. Additional baseline models and more details are demonstrated in Appendix A.2.

4.2 Results

We first compare the baselines with LigoSpace, which integrates GeoREC, Union-Pocket within the GNN, and the hybrid loss function on the DTIGN dataset. As shown in Table 1, our proposed method consistently enhances the DTIGN baseline across all subsets and metrics, demonstrating robust improvements on diverse protein targets. On average, our proposed method reduces RMSE by 9.93%, increases r by 18.39%, and improves τ by 19.09%. On the GIGN and GAT baselines, our method also achieves noticeable gains on 7 out of 8 and 6 out of 8 datasets, respectively. These results highlight the consistent effectiveness of our proposed method. Table 1 also shows that the proposed method achieves more improvement on stronger baselines.

Furthermore, we evaluate the performance of our proposed method on the SIU dataset, which includes a broader set of protein targets and ligands. The results are presented in Table 2, with the Uni-Mol [Zhou et al., 2023] benchmark taken from [Huang et al., 2025]. The results demonstrate that our LigoSpace-enhanced models outperform their respective baselines in 43 out of 48 dataset-metric combinations and achieve state-of-the-art performance. Notably, our method achieves substantial gains in some subsets. For example, in SIU-0.9/K_{i}, DTIGN+LigoSpace achieves $r = 0.3213$ and Spearman correlation of 0.3259. In contrast, the baseline methods struggle to learn meaningful relationships, as reflected by near-zero correlation scores. These results demonstrate the superior capability of our method on more complex datasets, further illustrating its robustness and generalizability. Additional experimental results and comparisons with more baseline models are provided in Appendix A.4.

4.3 Ablation study

To assess the contribution of each individual component in our proposed method, we perform an ablation study by systematically removing one module at a time from the complete configuration on the DTIGN model. The setups are as follows: 1) **DTIGN-GUL** includes GeoREC (G), Union-Pocket (U), and hybrid loss (L); 2) **DTIGN-GU(L)** includes GeoREC (G) and Union-Pocket (U), excluding the hybrid loss (L); 3) **DTIGN-(G)UL** includes the Union-Pocket (U) and hybrid loss (L), excluding GeoREC (G), and 4) **DTIGN-G(U)L** includes GeoREC (G) and hybrid loss (L), excluding the Union-Pocket (U). The results of the ablation study are shown in Table 3. By comparing these ablation settings with the full enhanced method (DTIGN-GUL), we draw the following conclusions:

The proposed geometric feature GeoREC (G) makes a substantial contribution to performance. When comparing DTIGN-(G)UL to DTIGN-GUL, we observe a reduction in RMSE by 9.3%, an increase in r by 14.4%, and an elevation in τ by 19%.

Table 1: Comparison of model performance on baselines and their LigoSpace-enhanced versions with the proposed methods across multiple datasets

Model	Metric	Dataset									
		I1	I2	I3	I4	I5	E1	E2	E3	Avg	Avg Imp%
DTIGN baseline	RMSE ↓	1.1977	0.7952	1.1443	0.9396	0.9518	0.9086	0.7765	1.0869	0.9751	-
	r ↑	0.3547	0.7128	0.7839	0.6177	0.6227	0.3363	0.6946	0.4825	0.5757	-
	τ ↑	0.2445	0.4922	0.5991	0.4574	0.4691	0.2139	0.4917	0.3786	0.4183	-
DTIGN+LigoSpace	RMSE ↓	**1.0364**	**0.6507**	**1.0245**	**0.8124**	**0.8966**	**0.8645**	**0.7262**	**1.0150**	**0.8783**	9.93%
	r ↑	**0.5895**	**0.7888**	**0.8251**	**0.7334**	**0.6986**	**0.4969**	**0.7275**	**0.5924**	**0.6815**	18.39%
	τ ↑	**0.4239**	**0.5454**	**0.6360**	**0.5205**	**0.5083**	**0.3705**	**0.5243**	**0.4563**	**0.4982**	19.09%
GIGN baseline	RMSE ↓	1.2365	0.7487	1.0677	0.8560	0.8700	0.9563	0.7545	0.9704	0.9325	-
	r ↑	0.4357	0.7234	0.8080	0.6888	0.7302	0.4435	0.7246	0.6285	0.6478	-
	τ ↑	0.3216	0.5083	0.6310	0.4920	0.5439	0.3443	0.5043	0.4664	0.4765	-
GIGN+LigoSpace	RMSE ↓	**0.9515**	**0.6880**	**0.9403**	**0.8111**	**0.8302**	**0.8733**	0.8737	**0.9161**	**0.8605**	7.72%
	r ↑	**0.6861**	**0.7735**	**0.8539**	**0.7450**	**0.7513**	**0.4912**	0.6637	**0.6860**	**0.7063**	9.03%
	τ ↑	**0.4899**	**0.5378**	**0.6673**	**0.5439**	**0.5591**	**0.3786**	**0.5511**	**0.5228**	**0.5313**	11.51%
GAT baseline	RMSE ↓	1.1801	0.9270	1.3433	1.0200	1.0547	0.8962	0.8579	1.1716	1.0564	-
	r ↑	0.4147	0.4771	0.6663	0.5091	0.5035	0.3771	0.5836	0.3468	0.4848	-
	τ ↑	0.2267	0.3115	0.4618	0.3685	0.3428	0.2451	0.3917	0.2938	0.3302	-
GAT+LigoSpace	RMSE ↓	**1.1767**	**0.8906**	**1.3352**	**1.0037**	**1.0342**	0.9252	0.8587	**1.1476**	**1.0465**	0.93%
	r ↑	**0.4511**	**0.5437**	**0.6813**	**0.5570**	**0.5302**	0.3604	**0.5838**	**0.4060**	**0.5142**	6.07%
	τ ↑	**0.2922**	**0.3593**	**0.4823**	**0.4075**	**0.3690**	**0.2497**	0.3819	**0.3330**	**0.3594**	8.82%

Table 2: Comparison of model performance on SIU dataset

Dataset / label	Method	RMSE↓	MAE↓	r↑	Spearman↑
SIU-0.9 / K_{d}	Unimol	1.364	1.141	-0.033	-0.082
	DTIGN	1.839	1.490	-0.001	-0.042
	DTIGN+LigoSpace	**1.304**	**1.060**	**0.321**	**0.326**
	GIGN	1.708	1.367	0.070	0.038
	GIGN+LigoSpace	**1.455**	**1.139**	**0.296**	**0.261**
	GAT	1.545	1.240	0.092	0.082
	GAT+LigoSpace	**1.473**	**1.166**	**0.261**	**0.254**
SIU-0.9 / K_{i}	Unimol	1.235	1.017	0.485	0.452
	DTIGN	1.607	1.276	0.360	0.329
	DTIGN+LigoSpace	**1.296**	**1.054**	**0.485**	**0.441**
	GIGN	1.597	1.337	0.223	0.167
	GIGN+LigoSpace	**1.487**	**1.240**	**0.371**	**0.338**
	GAT	1.706	1.386	0.301	0.262
	GAT+LigoSpace	**1.625**	**1.339**	**0.316**	**0.294**
SIU-0.6 / K_{d}	Unimol	1.389	1.192	-0.149	-0.206
	GIGN	1.371	1.115	0.265	0.281
	GIGN+LigoSpace	**1.326**	**1.078**	**0.280**	0.227
	DTIGN	1.349	1.079	0.329	0.329
	DTIGN+LigoSpace	**1.332**	**1.069**	0.327	0.248
	GAT	1.521	1.285	0.233	0.157
	GAT+LigoSpace	**1.424**	**1.182**	0.107	0.133
SIU-0.6 / K_{i}	Unimol	1.255	1.034	0.472	0.452
	GIGN	1.789	1.503	0.225	0.230
	GIGN+LigoSpace	**1.404**	**1.165**	**0.498**	**0.463**
	DTIGN	1.993	1.653	0.123	0.091
	DTIGN+LigoSpace	**1.321**	**1.079**	**0.472**	**0.452**
	GAT	1.976	1.690	-0.060	-0.096
	GAT+LigoSpace	**1.694**	**1.381**	**0.303**	**0.263**

The Union-Pocket also confers consistent advantages in the majority of cases. Compared to DTIGN-G(U)L, which excludes the Union-Pocket (U), it reduces the RMSE by 5.8%, increases r by 7%, and improves τ by 9% on average. Notably, GeoREC and Union-Pocket sometimes exhibit a mutually reinforcing relationship; when utilized in isolation, their influence may be restricted. For instance, on

 https://doi.org/10.52202/085713-3543

dataset E1, DTIGN-(G)UL and DTIGN-G(U)L yield r of approximately 0.31 and 0.32, whereas the DTIGN-GUL boosts r to 0.5, marking a nearly 60% improvement. This demonstrates that excluding either GeoREC (G) or Union-Pocket (U) results in substantial performance degradation, whereas their integration leads to a reinforced effect.

Integrating the hybrid loss (L) enhances the training process, particularly in terms of order consistency. This is evidenced by comparing DTIGN-GUL with DTIGN-GU(L); across 8 datasets, τ improves in 7 cases. On average, compared to DTIGN-GU(L), it reduces the RMSE by 3.1%, increases r by 4%, and improves τ by 4.6%. This validates the effectiveness of incorporating the pairwise loss in bioactivity prediction.

Table 3: Ablation study on DTIGN baseline (each variant excludes one component)

Dataset	DTIGN-GUL			DTIGN-GU(L)			DTIGN-(G)UL			DTIGN-G(U)L		
	RMSE↓	r↑	τ↑	RMSE↓	r↑	τ↑	RMSE↓	r↑	τ↑	RMSE↓	r↑	τ↑
I1	**1.0364**	0.5895	**0.4239**	1.1163	0.5259	0.3494	1.1341	0.4721	0.3249	1.0739	**0.5941**	0.4128
I2	**0.6507**	**0.7888**	**0.5454**	0.7292	0.7284	0.5081	0.7724	0.7077	0.4665	0.7395	0.7331	0.5265
I3	**1.0245**	**0.8251**	**0.6360**	1.0557	0.8153	0.6175	1.1170	0.8196	0.6348	1.2061	0.7486	0.5446
I4	**0.8124**	**0.7334**	**0.5205**	0.8550	0.7072	0.5023	0.9149	0.6666	0.4725	0.8364	0.7149	0.5067
I5	0.8966	0.6986	0.5083	**0.8625**	**0.7127**	**0.5225**	0.9591	0.6343	0.4561	0.9017	0.6928	0.4994
E1	0.8645	**0.4969**	**0.3705**	**0.8586**	0.4513	0.3498	0.9513	0.3064	0.1727	0.9243	0.3176	0.1983
E2	**0.7262**	0.7275	0.5243	0.7600	0.7083	0.5066	0.8263	0.6384	0.4300	0.7323	**0.7314**	**0.5332**
E3	1.0150	0.5924	**0.4563**	**1.0119**	**0.5955**	0.4524	1.0679	0.5198	0.3917	1.0438	0.5656	0.4332
Avg	**0.8783**	**0.6815**	**0.4982**	0.9062	0.6556	0.4761	0.9679	0.5956	0.4187	0.9323	0.6373	0.4568

5 Conclusion

Bioactivity prediction is a crucial task in computational drug discovery. In this work, we introduced GeoREC, a geometric representation of spatial emptiness in protein-ligand complexes, and the Union-Pocket, a unified structural context for multiple pockets, to enhance bioactivity prediction in drug discovery. By measuring unoccupied space and integrating global pocket information, our approach captures critical aspects of competitive binding that were previously overlooked. Additionally, we exploited a pairwise loss function to better preserve the relative order among samples. Extensive experiments across diverse datasets and bioactivity measurements demonstrate that LigoSpace significantly improves prediction accuracy. Specifically, the substantial performance gains achieved by GeoREC highlight the critical importance of the proposed geometric feature in protein-ligand interactions modeling.

Acknowledgments and Disclosure of Funding

This work was partially supported by Singapore Ministry of Education (MOE) Tier 1 RG97/22.

References

Antonio Baici. Kinetics of enzyme-modifier interactions. *Selected Topics in the Theory and Diagnosis of Inhibition and Activation Mechanisms*, 1, 2015.

Pedro J. Ballester and John B. O. Mitchell. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinform.*, 2010.

Elena L Cáceres, Nicholas C Mew, and Michael J Keiser. Adding stochastic negative examples into machine learning improves molecular bioactivity prediction. *Journal of Chemical Information and Modeling*, 60(12):5957–5970, 2020.

Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250, 2018.

Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, et al. Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, 57(12): 4977–5010, 2014.

Hyone-Myong Eun. *Enzymology primer for recombinant DNA technology*. Elsevier, 1996.

Daniel Fernández-Llaneza, Silas Ulander, Dea Gogishvili, Eva Nittinger, Hongtao Zhao, and Christian Tyrchan. Siamese recurrent neural network with a self-attention mechanism for bioactivity prediction. *ACS omega*, 6(16):11086–11094, 2021.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Holger Gohlke, Manfred Hendlich, and Gerhard Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *Journal of molecular biology*, 2000.

Jiyun Han, Shizhuo Zhang, Mingming Guan, Qiuyu Li, Xin Gao, and Juntao Liu. Geonet enables the accurate prediction of protein-ligand binding sites through interpretable geometric deep learning. *Structure*, 32(12):2435–2448, 2024.

Yanwen Huang, Bowen Gao, Yinjun Jia, Hongbo Ma, Wei-Ying Ma, Ya-Qin Zhang, and Yanyan Lan. Redefining the task of bioactivity prediction. In *The Thirteenth International Conference on Learning Representations*, 2025.

Muhammad Ishfaq, Muhammad Aamir, Farooq Ahmad, Abdelazim M Mebed, and Sayed Elshahat. Machine learning-assisted prediction of the biological activity of aromatase inhibitors and data mining to explore similar compounds. *ACS omega*, 7(51):48139–48149, 2022.

IM2443682 Kapetanovic. Computer-aided drug discovery and development (caddd): in silico-chemico-biological approach. *Chemico-biological interactions*, 171(2):165–176, 2008.

Sarah L. Kinnings, Nina Liu, Peter J. Tonge, Richard M. Jackson, Lei Xie, and Philip E. Bourne. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J. Chem. Inf. Model.*, 2011.

Alpha A Lee, Qingyi Yang, Asser Bassyouni, Christopher R Butler, Xinjun Hou, Stephen Jenkinson, and David A Price. Ligand biological activity predicted by cleaning positive and negative chemical correlations. *Proceedings of the National Academy of Sciences*, 116(9):3373–3378, 2019.

Shuangli Li, Jingbo Zhou, Tong Xu, Liang Huang, Fan Wang, Haoyi Xiong, Weili Huang, Dejing Dou, and Hui Xiong. Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 975–985, 2021.

Eric MarÉchal. Measuring bioactivity: Ki, ic50 and ec50. *Chemogenomics and Chemical Genetics: A User's Introduction for Biologists, Chemists and Informaticians*, pages 55–65, 2011.

Andrea Mastropietro, Giuseppe Pasculli, and Jürgen Bajorath. Learning characteristics of graph neural networks predicting protein–ligand affinities. *Nature Machine Intelligence*, 5(12):1427–1436, 2023.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019.

Hakime Öztürk, Arzucan Özgür, and Elif Özkirimli Ölmez. Deepdta: deep drug-target binding affinity prediction. *Bioinform.*, 2018.

Parixit Prajapati, Princy Shrivastav, Jigna Prajapati, and Bhupendra Prajapati. Deep learning approaches for predicting bioactivity of natural compounds. *The Natural Products Journal*, 2025.

Bharath Ramsundar, Peter Eastman, Pat Walters, and Vijay Pande. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. O'Reilly Media, 2019.

Ahmet Sureyya Rifaioglu, Heval Atas, Maria Jesus Martin, Rengul Cetin-Atalay, Volkan Atalay, and Tunca Doğan. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in bioinformatics*, 20(5):1878–1912, 2019.

 https://doi.org/10.52202/085713-3543

Bonggun Shin, Sungsoo Park, Keunsoo Kang, and Joyce C Ho. Self-attention based molecule representation for predicting drug-target interaction. In *Machine learning for healthcare conference*. PMLR, 2019.

Zhenqiao Song, Tinglin Huang, Lei Li, and Wengong Jin. Surfpro: Functional protein design based on continuous surface. *arXiv preprint arXiv:2405.06693*, 2024.

Toshitaka Tanebe and Takashi Ishida. End-to-end learning for compound activity prediction based on binding pocket information. *BMC bioinformatics*, 22:1–11, 2021.

Alexander Tropsha, Olexandr Isayev, Alexandre Varnek, Gisbert Schneider, and Artem Cherkasov. Integrating qsar modelling and deep learning in drug discovery: the emergence of deep qsar. *Nature Reviews Drug Discovery*, 23(2):141–155, 2024.

Michael Tynes, Wenhao Gao, Daniel J Burrill, Enrique R Batista, Danny Perez, Ping Yang, and Nicholas Lubbers. Pairwise difference regression: a machine learning meta-algorithm for improved prediction and uncertainty quantification in chemical search. *Journal of chemical information and modeling*, 61(8):3846–3857, 2021.

Ali Vefghi, Zahed Rahmati, and Mohammad Akbari. Drug-target interaction/affinity prediction: Deep learning models and advances review. *arXiv preprint arXiv:2502.15346*, 2025.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Izhar Wallach, Michael Dzamba, and Abraham Heifets. Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*, 2015.

Renxiao Wang, Luhua Lai, and Shaomeng Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.*, 2002.

Zechen Wang, Sheng Wang, Yangyang Li, Jingjing Guo, Yanjie Wei, Yuguang Mu, Liangzhen Zheng, and Weifeng Li. A new paradigm for applying deep learning to protein–ligand interaction prediction. *Briefings in bioinformatics*, 25(3):bbae145, 2024.

Zhewei Wei, Ye Yuan, Chongxuan Li, Wenbing Huang, et al. Equipocket: an e (3)-equivariant geometric graph neural network for ligand binding site prediction. In *Forty-first International Conference on Machine Learning*, 2024.

Ziduo Yang, Weihe Zhong, Qiujie Lv, Tiejun Dong, and Calvin Yu-Chian Chen. Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3d structures (gign). *The journal of physical chemistry letters*, 2023.

Yueming Yin, Haifeng Hu, Zhen Yang, Huajian Xu, and Jiansheng Wu. Realvs: toward enhancing the precision of top hits in ligand-based virtual screening of drug leads from large compound databases. *Journal of chemical information and modeling*, 61(10):4924–4939, 2021.

Yueming Yin, Haifeng Hu, Zhen Yang, Feihu Jiang, Yihe Huang, and Jiansheng Wu. Afse: towards improving model generalization of deep graph learning of ligand bioactivities targeting gpcr proteins. *Briefings in Bioinformatics*, 23(3):bbac077, 2022.

Yueming Yin, Hilbert Yuen In Lam, Yuguang Mu, Hoi-Yeung Li, and Adams Wai-Kin Kong. Advancing bioactivity prediction through molecular docking and self-attention. *IEEE J. Biomed. Health Informatics*, 2024.

Xiangying Zhang, Haotian Gao, Haojie Wang, Zhihang Chen, Zhe Zhang, Xinchong Chen, Yan Li, Yifei Qi, and Renxiao Wang. Planet: a multi-objective graph neural network model for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling*, 64(7):2205–2220, 2023.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Vijil, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. In *International Conference on Machine Learning*, 2022.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.

Dixian Zhu, Tianbao Yang, and Livnat Jerby. Gradient aligned regression via pairwise losses. *arXiv preprint arXiv:2402.06104*, 2024.

https://doi.org/10.52202/085713-3543

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarized the key contributions at the end of the introduction. They can accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Due to space constraints in the main text, we discuss the limitations of this work in the "Limitations" section of the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Detailed instructions for reproducing our experimental results are provided in the "Implementation" section of the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/085713-3543

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: At this time, we are not able to share the code publicly, as a patent application is in progress and there may be future commercial considerations.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Necessary experimental settings are described in Section 4.1 of the main text. Additional details are provided in the "Detailed experimental settings" section of the Appendix due to space constraints.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported because it would be too computationally expensive.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided the computer resources needed to reproduce the experiments in the "Computational environment" section of the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and made sure to adhere to them.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our method improves bioactivity prediction, which can help accelerate drug discovery, as outlined in the Introduction and Related Work sections. We are not aware of any potential negative societal impacts at this time.

https://doi.org/10.52202/085713-3543

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly respect and credit for existing assets. We cite relevant papers and comply with the licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

 https://doi.org/10.52202/085713-3543

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-3543